中北大学学报(自然科学版)
JOURNAL OF NORTH UNIVERSITY OF CHINA(NATURAL SCIENCE EDITION)

文章编号：1673-3193(XXXX)XX-0001-12

基于深度强化学习的智能驾驶方法研究与部署实现

杨双龙，罗　佳*，贾子厚，刘正阳
（中北大学 能源与动力工程学院，山西 太原 030051）

摘　要：深度强化学习在自动驾驶领域展现出巨大潜力，但仍普遍存在学习效率低、动作连续性差以及从仿真到实车的迁移部署困难等问题。为此，本文提出一种融合预训练变分自编码器(Variational Autoencoder，VAE)、长短期记忆网络(Long Short-Term Memory，LSTM)与近端策略优化(Proximal Policy Optimization，PPO)的智能驾驶方法，构建了LSTM-PPO算法框架。该方法利用VAE从语义分割图像中提取95维环境特征，并与5维车辆状态向量融合，作为策略网络的输入；通过统一的编码-解码结构与训练阶段注入噪声，实现了从CARLA仿真环境到实车的零样本Sim-to-Real迁移。实验结果表明：在仿真训练中，LSTM-PPO将累计碰撞次数从802次降低至208次，且轨迹偏移收敛更快；实车测试中，直线行驶的平均横向偏移由5.436 cm降至2.268 cm，整圈行驶的偏移量均方根误差由19.868 cm降至7.356 cm，任务成功率从35.71%提升至62.50%。仿真与实车部署结果共同验证了所提方法在提升轨迹平滑性、控制稳定性与跨域部署鲁棒性方面的显著优势。

关键词：自动驾驶；深度强化学习；决策控制；近端策略优化；实车部署
中图分类号：U46　　**文献标识码：**A　　**doi:** 10.62756/jnuc.issn.1673-3193.2025.07.0012
引用格式：杨双龙,罗佳,贾子厚等.基于深度强化学习的智能驾驶方法研究与部署实现[J].中北大学学报(自然科学版),DOI:10.62756/jnuc.issn.1673-3193.2025.07.0012.
Yang Shuanglong, Luo Jia, Jia Zihou, et al. Research and Deployment of Intelligent Driving Method Based on Deep Reinforcement Learning [J]. Journal of North University of China (Natural Science Edition), DOI: 10.62756/jnuc.issn.1673-3193.2025.07.0012.

Research and Deployment of Intelligent Driving Method Based on Deep Reinforcement Learning

Yang Shuanglong, Luo Jia*, Jia Zihou, Liu Zhengyang
(School of Energy and Power Engineering, North University of China, Taiyuan 030051, China)

Abstract: Deep reinforcement learning (DRL) has shown considerable potential in autonomous driving, yet it still faces challenges such as low learning efficiency, poor action continuity, and difficulty in sim-to-real transfer and deployment. To address these issues, this paper proposed an intelligent driving method that integrated a pre-trained variational autoencoder (VAE), a long short-term memory (LSTM) network, and the proximal policy optimization (PPO) algorithm, forming an LSTM-PPO framework. The method employed the VAE to extract a 95-dimensional environmental feature vector from semantic seg-

收稿日期：2025-07-23
基金项目：校企合作项目(2412000072HX)
作者简介：杨双龙(2000—)，男，硕士生，主要从事深度学习、智能汽车决策控制技术等研究。
*** 作者简介：**罗　佳(1981—)，女，讲师，博士，主要从事智能汽车环境感知、智能控制技术及应用方面的研究。E-mail:sjj1314@nuc.edu.cn。

mentation images, which was then fused with a 5-dimensional vehicle state vector as input to the policy network. By adopting a unified encoder-decoder structure and injecting noise during training, zero-shot Sim-to-Real transfer from the CARLA simulator to a physical vehicle was achieved. Experimental results show that during simulation training, LSTM-PPO reduces the cumulative number of collisions from 802 to 208 and accelerates the convergence of trajectory deviation. In real-vehicle tests, the average lateral offset in straight-line driving decreases from 5.436 cm to 2.268 cm, while in full-lap driving, the root-mean-square error of the lateral offset drops from 19.868 cm to 7.356 cm, and the success rate increases from 35.71% to 62.50%. Both simulation and real-world deployment outcomes demonstrate the significant advantages of the proposed method in improving trajectory smoothness, enhancing motion stability, and strengthening deployment robustness.

Key words: autonomous driving; deep reinforcement learning; decision-making and control; proximal policy optimization; real-world vehicle deployment

0 引 言

随着人工智能技术的快速发展，自动驾驶技术正逐步从实验室走向实际应用。作为自动驾驶的核心技术之一，深度强化学习(Deep Reinforcement Learning,DRL)因其强大的非线性建模能力和端到端的学习框架，成为智能车辆自主驾驶研究的重要方向[1-3]，并在自动驾驶领域的研究取得了显著进展[4-6]。Zhang等[7]提出了一种基于分布式强化学习的轨迹规划算法，通过利用RL(Reinforcement Learning)生成多模式轨迹。显著提升了复杂城市交通场景下的路径规划效率和安全性。这一方法在nuPlan大规模真实数据集上验证了其超越传统规则驱动方案的性能。在端到端方面，Zhao等[8]提出了一种名为CAscade Deep REinforcement learning framework(CADRE)的级联深度强化学习框架，用于基于视觉的城市自动驾驶。该方法通过联合训练协同注意力感知模块和强化学习策略，实现了在复杂城市环境中的高效驾驶。

针对强化学习中的探索-利用平衡问题，Hafner等[9]提出了DreamerV3算法，通过世界模型(World Model)构建虚拟环境进行离线预训练，结合在线强化学习策略优化，大幅降低了真实路测数据需求，并提升了长尾场景的泛化能力。该算法已在英伟达DRIVE平台上实现了商业化部署。在时序决策建模方面，Chen等[10]提出了一种可解释的端到端深度强化学习方法。该方法引入了序列隐环境模型，能够生成语义鸟瞰图，提供对驾驶策略的可解释性，并在城市驾驶场景中展示了优越的性能。

尽管已有研究在一定程度上推动了自动驾驶技术的发展，但当前方法仍面临若干核心挑战[11-14]：1)现有方法多将环境感知与规划决策整合为单一任务进行训练，这种方式易导致学习效率偏低、模型收敛稳定性不足，且训练过程耗时较长。2)强化学习中采用的随机探索策略往往难以确保动作输出的连续性，致使车辆在直线行驶时出现左右摆动，进而影响驾驶的平稳性与安全性。3)多数现有研究尚未通过实际部署验证方法的有效性和实用性。

为应对上述不足，本文提出了一种融合预训练VAE与LSTM-PPO的自动驾驶决策框架。具体而言，通过预训练变分自编码器(VAE)对语义分割图像进行特征提取，将提取的分布特征与车辆状态信息相融合，作为强化学习模型的输入，从而显著提升算法的学习效率并增强模型收敛的可靠性。同时，本研究设计了基于LSTM-PPO的决策模型，利用长短时记忆网络(LSTM)捕捉训练过程中的时序上下文特征，赋予智能体一定的预测能力，以优化决策过程的连续性和准确性。

1 相关工作

1.1 强化学习

强化学习的基本框架由智能体(Agent)、环境(Environment)、状态(State)、动作(Action)以及奖励(Reward)五个核心要素构成[15](如图1所示)。在这一机制下，智能体通过执行特定动作促使环境发生状态转移，随即环境根据新的状态生成相应的奖励信号(可为正向激励或负向反馈)。智能体基于此新状态及奖励信息，按照既定策略选择后续动作。这一循环过程体现了智能体与环境通过状态、动作和奖励的动态交互模式。通过

强化学习的迭代训练，智能体逐步优化其决策能力，以适应复杂多变的环境需求。

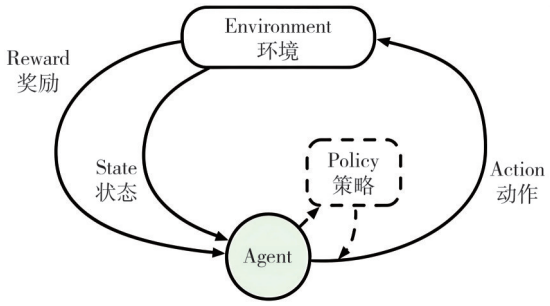


图1 强化学习组成图

Fig. 1 Block diagram of reinforcement learning components

1.2 自动变分编码器

变分自动编码器(VAE)是一种基于生成范式的神经网络架构(如图 2 所示)，由编码器和解码器两大模块构成。编码器将输入x转换为潜在变量z的高斯分布参数(均值u和方差σ^2)，而解码器则从潜在变量$z=u+\sigma*\varepsilon$(其中$\varepsilon\sim N(0,1)$)重构原始数据。其核心机制在于将高维输入数据映射至低维潜在空间，随后通过解码器重构回高维形式[16-18]。通过量化重构输出与原始输入间的偏差，网络参数得以利用反向传播算法进行精细调整。VAE的编码器与解码器均为非线性变换函数，通常依托多层感知器(MLP)或卷积神经网络实现，并具备通过梯度优化迭代提升性能的能力。

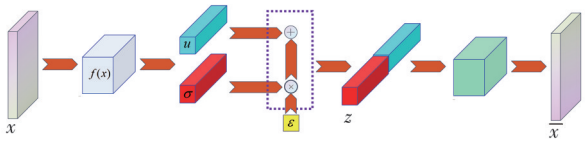


图 2 VAE结构图

Fig. 2 Schematic of the VAE architecture

1.3 长短期记忆网络

长短期记忆网络(LSTM)是对传统循环神经网络(Recurrent Neural Network,RNN)的改进，如图 3 所示，其设计旨在解决梯度消失问题，特别是在处理长序列时的依赖问题。LSTM通过引入三个关键门控机制：遗忘门f_t、输入门I_t和输出门O_t，使得网络能够在每个时间步对信息进行选择性地存储、遗忘和输出[19]。这些门控机制通过控制信息的流动，确保网络能够有效地保存和更新长期记忆，同时避免过多无关信息的干扰。

遗忘门主要利用sigmoid函数，决定上一时刻网络的输出h_{t-1}和上一时刻网络的单元状态C_{t-1}是否继续存在于当前时刻网络的单元状态C_t中[20]。遗忘门计算公式为

$$f_t=\sigma(\boldsymbol{W}_f\cdot g[h_{t-1},x_t]+b_f), \tag{1}$$

式中：$\boldsymbol{W}_f$为权值矩阵；b_f为偏置量；x_t为当前网络的输入；g为向量拼接。

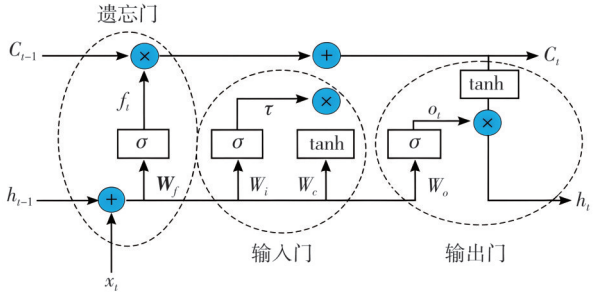


图 3 LSTM网络结构图

Fig. 3 Schematic of the LSTM network architecture

输入门利用sigmoid函数输出的信息与tanh函数输出的信息相乘，决定当前时刻的输入x_t有多少要传到单元状态C_t中。输入门计算公式为

$$i_t=\sigma(W_i\cdot g[h_{t-1},x_1]+b_i)\cdot\tanh(W_c\cdot g[h_{t-1},x_t]+b_c)。 \tag{2}$$

输出门也是利用sigmoid函数与tanh函数输出的信息相乘，决定单元状态C_t中有多少可以传到当前输出h_t中。输出门的计算公式为

$$h_t=\sigma(W_0\cdot g[h_{t-1},x_t]+b_0)\cdot\tanh(C_t)。 \tag{3}$$

2 基于LSTM-PPO算法的智能驾驶方法

2.1 系统整体设计

本文整体框架如图 4 所示，采用“模拟训练-真实部署”的双阶段架构设计。左侧紫色虚线框内为基于CARLA模拟器的训练过程，右侧红色虚线框内为本地智能小车的实际部署过程。

训练阶段通过CARLA仿真平台构建高保真训练环境。模型采用多模态特征融合机制：首先通过语义分割网络提取道路场景的像素级语义特征，利用编码器将其压缩为低维特征向量；同时整合车辆状态数据(包括速度、转向角、加速度等动态参数)与环境特征，形成复合状态表征。该状态空间输入到基于PPO算法搭建的智能体进行训练优化，最终输出连续动作空间a(包含转向角、油门与刹车控制量)，完成端到端的驾驶策略学习。

部署阶段通过模型直接迁移实现仿真到现实(Sim-to-Real)的跨域适配。智能小车使用搭载的摄像头实时采集道路图像，经轻量化语义分割模型处理后，采用与训练阶段同构的编码器进行特征降维。实时获取的车辆状态数据与编码后的环境特征融合后，输入已训练的强化学习策略网络。系统通过强化学习模型推理生成控制指令，形成“感知-决策-执行”的闭环控制流。

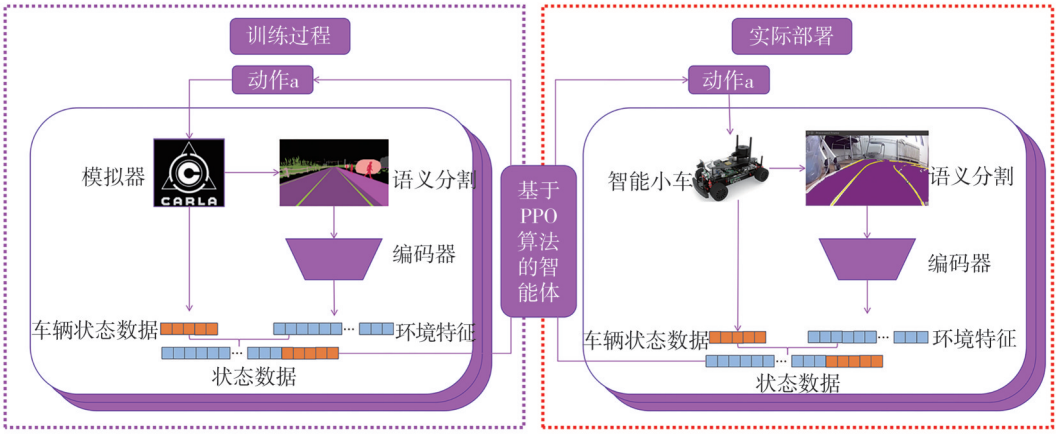


图 4　整体框架图

Fig. 4　Overall system architecture

2.2　预训练自动变分编码器

为了缓解高维图像状态空间所引发的强化学习模型训练复杂度问题，本文采用了一种预训练的自动变分编码器(VAE)模型，对获取到的分割图像进行特征降维。该方法通过对输入图像进行有效的编码，从而将高维图像数据映射至低维潜在空间，以减少计算开销并提高训练效率。本文所采用的VAE结构如图 5 所示。

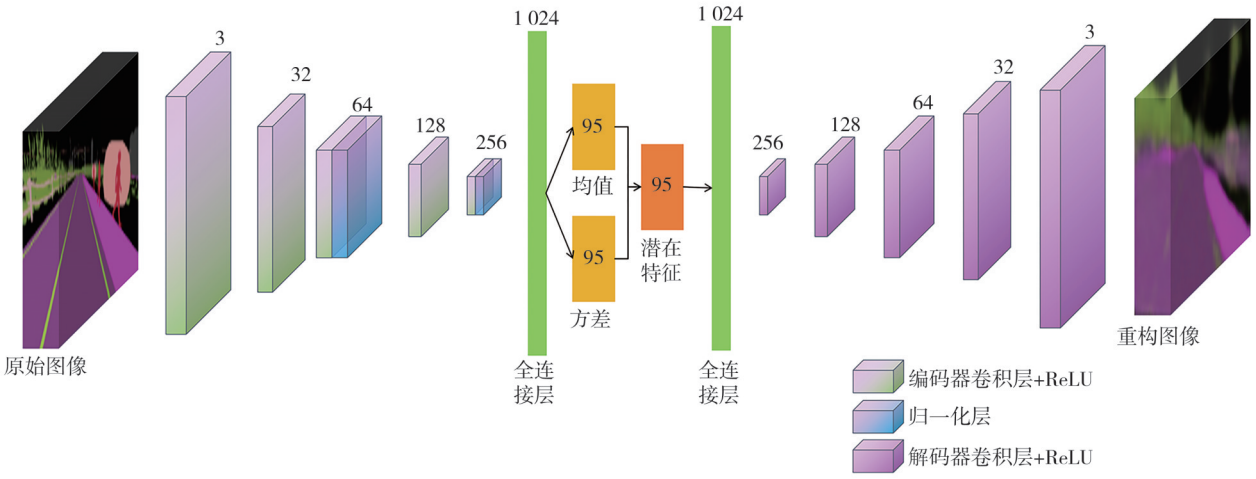


图 5　本文所用VAE结构图

Fig. 5　Schematic of the VAE used in this study

自动变分编码器(VAE)的训练过程始于通过自动驾驶与手动驾驶相结合的方式，在CARLA中收集了12 000张分辨率为160×80的语义分割图像，并将其展开为$h\times w\times c$=38 400维的输入向量馈入网络。值得注意的是，为了保证训练样本的多样性与代表性，在CARLA仿真环境中精心设计了多种道路与气象场景：包括市区道路(占样本总数的约40%)、乡村道路(约30%)、高速公路(约20%)三类典型道路类型；同时分别在晴朗、阴天、雨天和晨雾四种气象条件下各采集相同比例的图像，以模拟不同能见度与路面质感变化对语义分割精度的影响。在完成全部训练迭代后，VAE的网络权重即被冻结，以确保后续下游任务的输入空间保持一致。

2.3 LSTM-PPO强化学习模型

2.3.1 状态空间与动作空间

在深度强化学习框架下，状态空间的设计需全面整合智能小车对外部环境的感知能力及其自身的实时状态属性，例如车速与方向盘转角等。本研究的状态空间构造如图6所示，其中第一部分利用变分自动编码器(VAE)对图像数据进行压缩编码，生成95维的特征向量；第二部分则包含5个车辆状态参数，其中第1维至第3维分别对应车速、方向盘转角以及油门/制动踏板的量化描述。第4维表示车辆当前是否位于可行驶车道内，第5维表示车辆相对于车道中心线的横向偏移量。这些特征共同构成了智能小车在特定时刻的完整状态描述。

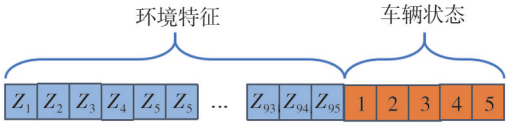


图6 状态空间示意图

Fig. 6 Schematic of the state space

本文的动作空间设计输出为

$$\varepsilon\in[-1,1]。\tag{4}$$

ε代表归一化后的方向盘转角，$\varepsilon>0$时，控制小车右转，反之左转。

$$\mu\in[-1,1]。\tag{5}$$

μ代表归一化后的油门/制动踏板开度，$\mu>0$时，控制油门踏板开度，反之控制刹车踏板开度。

2.3.2 奖励函数设计

在强化学习体系中，奖励函数的构建设计直接决定了模型参数优化的效能。稀疏的奖励设置常使模型收敛过程受阻，而过于密集的奖励则易使算法陷入局部最优困境。在城市道路自动驾驶场景下，奖励函数的制定需综合考量若干关键因素：首先，车辆需维持适宜的速度并严格遵循车道线行驶；其次，车辆前进方向与道路基准方向的偏差角应尽可能缩小；最后，车辆应尽量保持在道路中央行驶，从而提高行驶安全性和效率。基于这些考虑，本文提出了如下的奖励函数设计。

1) 速度控制奖励。在城市道路驾驶任务中，车辆的速度应保持在一个合理的区间$[v_{\min},v_{\max}]$内，以保证安全性和驾驶效率。因此，车辆的速度v_{current}仅在其处于该区间时才会获得正向奖励。若车辆的速度超出该区间，将不给予奖励。该部分奖励函数可表示为

$$R_{\text{speed}}=\begin{cases}1 & \text{if } v_{\min}\leqslant v_{\text{current}}\leqslant v_{\max},\\ 0 & \text{if } v_{\text{current}}<v_{\min} \text{ or } v_{\text{current}}>v_{\max}。\end{cases}\tag{6}$$

此设计确保了智能驾驶系统在控制车速时不会偏离安全范围，有助于避免因过快或过慢行驶带来的风险。

2) 与道路方向夹角奖励。车辆应尽量保持与道路方向的一致性，以避免出现不必要的转向或行驶偏差。为此，定义车辆当前行驶方向与道路方向之间的夹角θ_{vehicle}和θ_{road}的差异，并通过负值奖励函数激励车辆尽量与道路方向保持一致。

$$R_{\text{direction}}=-|\theta_{\text{vehicle}}-\theta_{\text{road}}|。\tag{7}$$

该奖励函数确保车辆在行驶过程中尽可能与道路方向对齐，从而提高行驶的稳定性与效率。

3) 车辆位置奖励。该奖励是为了保证车辆行驶时尽量保持在车道的中心位置。因此，通过计算车辆当前位置x_{vehicle}与车道中心位置x_{center}之间的距离来设计奖励函数，即

$$R_{\text{center}}=-|x_{\text{vehicle}}-x_{\text{center}}|。\tag{8}$$

此设计促使车辆尽可能地保持在车道中心行驶，从而提高了行驶的安全性与准确性。

4) 碰撞惩罚。为了避免智能驾驶系统发生碰撞，必须设计一个显著的惩罚项，以抑制碰撞行为的发生。当车辆与障碍物或其他物体发生碰撞时，将根据碰撞检测系统输出显著的负奖励，以促使系统在训练过程中优先选择避碰策略。碰撞惩罚可表示为

$$R_{\text{collision}}=\begin{cases}-10 & \text{if collision occurs,}\\ 0 & \text{if no collision。}\end{cases}\tag{9}$$

此惩罚机制能够有效促进智能驾驶系统在面对障碍物时采取规避措施，从而提高驾驶安全性。

5) 综合奖励。最终的综合奖励函数由速度控制、方向对齐、车道居中和碰撞惩罚四部分构成。通过加权求和，可得综合奖励函数

$$R_{\text{total}}=w_1R_{\text{speed}}+w_2R_{\text{direction}}+w_3R_{\text{center}}+w_4R_{\text{collision}},\tag{10}$$

式中：w_1，w_2，w_3，w_4为各项奖励的权重系数，根据实际应用需求进行调节。该综合奖励函数能够在多目标优化的基础上引导智能驾驶系统在城市道路环境中学习到有效的驾驶策略。

2.3.3 LSTM-PPO算法

PPO作为一种依托Actor-Critic架构的强化学习核心算法，相较于传统策略梯度方法，其创新之处在于融入了置信域理论，并通过施加KL散度约束，

确保策略迭代后回报函数呈现单调递增的特性。本研究将长短期记忆网络(LSTM)与PPO算法相结合，构建了LSTM-PPO算法框架，如图7所示。

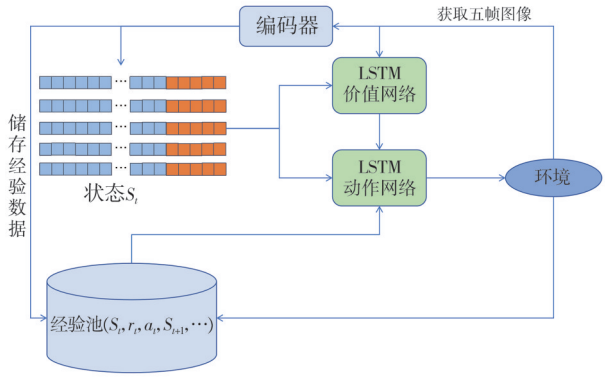


图7 LSTM-PPO算法结构图

Fig. 7 LSTM-PPO algorithm architecture

动作网络首先通过三层全连接神经网络对输入状态序列进行逐步降维与非线性变换：初始层将状态维度映射至500个隐藏单元，随后依次压缩至300个和100个单元，每层均辅以双曲正切(tanh)激活函数以增强非线性建模能力。随后，将全连接层输出的100维特征序列输入隐藏层维度为64的LSTM单元，通过其门控结构对连续观测状态之间的时间相关性进行建模，以提取更稳定的时序特征表示。LSTM的输出结果再经线性层映射至动作维度，并通过tanh激活函数约束动作均值范围。动作网络结构如图8所示。

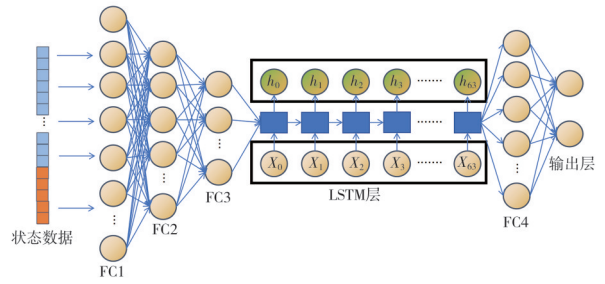


图8 动作网络结构图

Fig. 8 Architecture of the action network

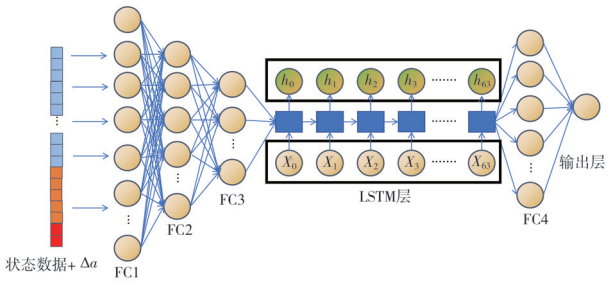


图9 价值网络结构

Fig. 9 Architecture of the value network

价值网络结构与动作网络结构类似，但输入数据由状态向量以及当前动作网络生成的控制动作的变化量组成。后续处理过程与动作网络相同。价值网络结构如图9所示。

3 Carla仿真及结果

3.1 Carla环境

本文算法的训练环境为Windows10，GPU为NVIDIA GeForce RTX 4080 SUPER，采用python3.7和torch=1.12.1+cu11.3。仿真环境使用Carla0.9.15中的Town02地图，如图10所示。为保证训练效率与训练结果的可解释性，每回合的实验终止条件设置如下：

1) 智能体与周围物体发生碰撞——用于检测控制策略的安全边界。

2) 智能体速度连续5 s小于2 km/h——由于采用静态交通环境训练，无交通拥堵干扰，该阈值主要用于捕捉控制失效或陷入停滞的情形。

3) 智能体速度连续3 s大于25 km/h——超出LSTM稳定控制预期范围的“失控”场景，以便及时终止并回收样本。

4) 智能体偏离车道中心线的距离大于1.5 m——考虑到Town02地图中车道宽度约为3 m，此阈值代表严重偏离单车道范围，超出稳定控制可接受的容忍度。

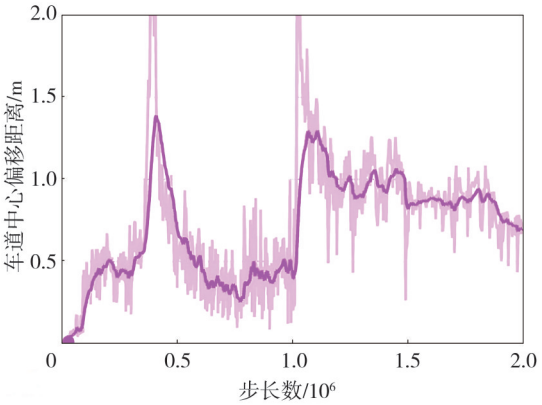
图10 Town02俯视图

Fig. 10 Top view of Town02

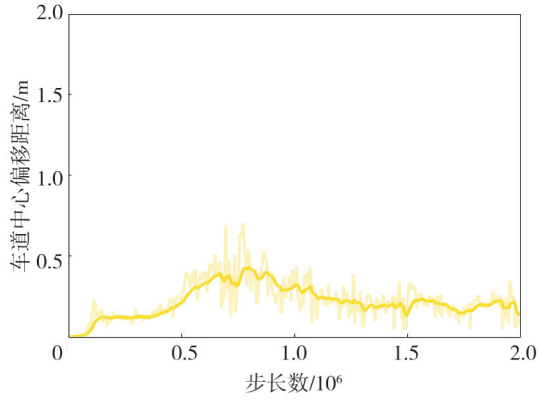
3.2 仿真训练结果

仿真训练结果如图11和图12所示。根据结果的分析，LSTM-PPO算法(如图11(b))在自动驾驶车道保持任务中相较于PPO算法(如图11(a))表现出显著的优越性。具体而言，从

车道中心偏移量(Lane Center Offset)的变化趋势来看，PPO算法的偏移量在训练过程中波动较大，峰值接近2 m，且始终维持较高的波动幅度，反映出车辆在行驶稳定性上的不足；而LSTM-PPO算法的偏移量虽在训练初期存在一定波动，但随后迅速收敛至较低水平，整体稳定在0.5 m以下，表明其在减少汽车左右摆动和提升车道保持能力方面具有明显优势。



(a) PPO模型



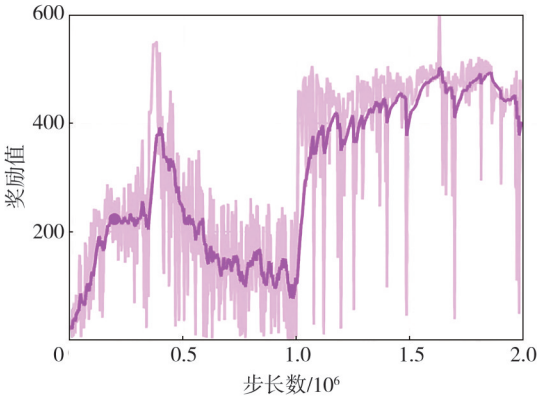
(b) LSTM-PPO模型

图11 道路中心偏移量

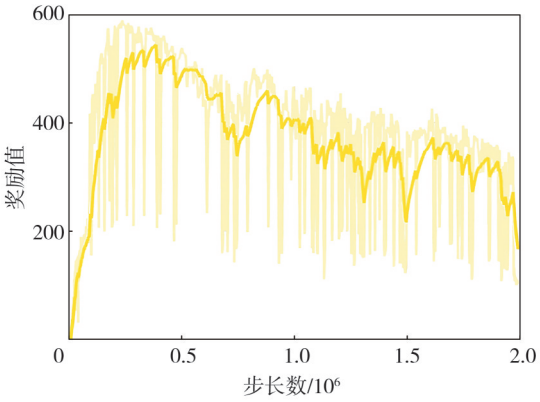
Fig. 11 Lateral offset from the road centerline

同时，从奖励值(Reward)的对比来看，LSTM-PPO算法的奖励值(图12(b))在训练初期快速提升，并持续保持较高水平，最高接近600，远超经典PPO算法(图12(a))的最高值(约450)，后者奖励值波动较大且整体趋势较低。这一结果进一步验证了LSTM-PPO算法在任务完成效率和训练稳定性上的优越性。

基于上述实验观测，进一步对累计碰撞次数进行分析(见图13)可知，LSTM-PPO模型相较于基础PPO模型在长期安全性上同样具有显著优势。具体表现为：在相同时间步长下，LSTM-PPO模型的碰撞累积量始终维持在较低水平，当仿真进行至1.4×10^6步时，其碰撞次数为127次，显著低于PPO模型的631次；至2×10^6步长仿真终止阶段，LSTM-PPO的累计碰撞量为208次，而PPO模型则为802次碰撞。值得注意的是，LSTM-PPO模型的碰撞增长曲线在超过0.6×10^6步长后呈现显著趋缓特征，最终趋于平稳，与PPO模型持续线性上升的趋势形成鲜明对比，这一现象进一步佐证了LSTM-PPO算法在长期避障稳定性上的优越性能。



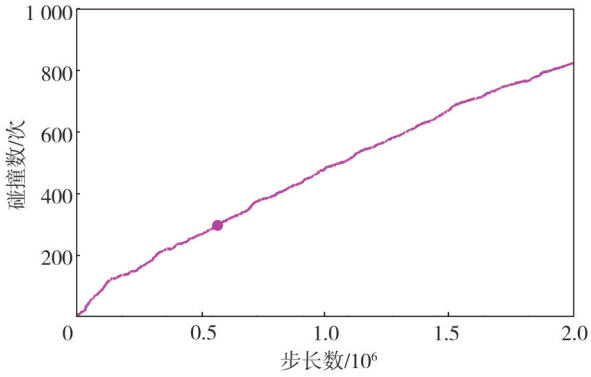
(a) PPO模型



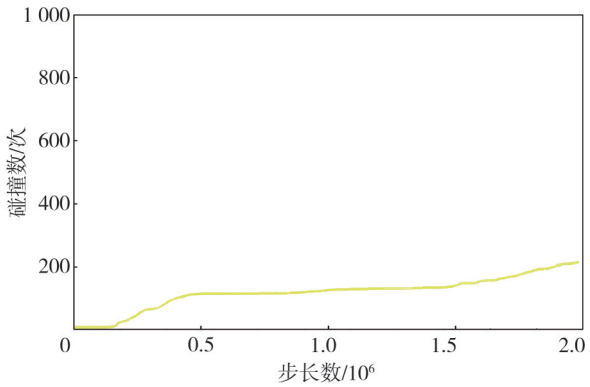
(b) LSTM-PPO模型

图12 奖励值图像

Fig. 12 Reward value plot



(a) PPO模型



(b) LSTM-PPO模型

图 13 累计碰撞次数图像

Fig. 13 Cumulative collision count plot

3.3 结果分析

LSTM-PPO算法通过引入长短期记忆网络(LSTM),有效捕捉了时间序列数据中的长期依赖关系,这一能力在自动驾驶场景中尤为关键,因为车辆运动状态与环境变化具有显著的时间相关性。相比之下,经典PPO算法主要依赖当前观测和短期历史信息,难以充分理解复杂动态环境中的长期模式,因而导致预测误差放大,进而表现为较大的车道中心偏移、较低的任务奖励和较高的碰撞次数。LSTM的门控机制使得LSTM-PPO能够记忆并利用历史观测数据,从而在训练过程中逐步优化策略,实现更精确的环境预测与稳定的决策输出。这种长期依赖关系的建模能力不仅降低了车辆行驶中的摆动幅度,提升了车道保持的稳定性,还通过更优的决策选择推动了奖励值的持续增长。因此,LSTM-PPO算法在实验中的优越性本质上源于其对时间维度信息的深度挖掘与利用,显著增强了算法在复杂任务环境下的适应性。

4 实车验证

4.1 实验平台

在部署阶段,本研究采用基于观测空间对齐与噪声建模的零样本Sim-to-Real迁移策略,而非在实车上进一步Fine-tuning。具体而言,训练与部署阶段均使用同一语义分割模型和相同的编码-解码结构,将CARLA仿真环境和真实环境中的相机图像统一映射为分辨率、类别集合及归一化方式一致的语义特征表示,并直接输入到仿真中训练完成的策略网络进行决策。为了进一步缩小仿真与现实在感知层面的差距,本研究在训练阶段对语义特征加入随机噪声进行数据增强,用以模拟实际传感器输出中的测量误差与感知抖动,从而提升策略对噪声干扰的鲁棒性。

在系统整体架构中,感知模块、计算平台(Jetson Orin NX)与底盘执行机构之间构成了一个闭环的数据流与控制信号流链路。具体而言,前端摄像头将采集到的原始道路图像通过高速接口(USB3.0)实时传输至Jetson Orin NX,在计算平台上首先由轻量化语义分割网络进行前向推理,生成仅保留车道线、道路边界及可行驶区域等关键信息的语义分割结果。该语义分割图像随后被送入与训练阶段保持一致的编码器进行特征提取与降维,输出低维环境特征向量。同时,底盘侧的传感器与控制单元(如里程计、IMU、编码器等)通过串口周期性上传车辆线速度、航向角、横向偏移等状态信息至Jetson Orin NX,二者在计算平台内被时序对齐并融合,构成LSTM-PPO策略网络的输入。策略网络在Jetson Orin NX上完成推理后输出转向角和目标速度控制量,这些控制指令经由通信接口发送至底盘控制器,由底盘控制器进一步转换为对驱动电机与转向舵机的具体驱动信号,从而实现对智能小车纵向与横向运动的精细控制。通过上述从"感知-决策-执行"的闭环数据流与控制信号流设计,可以确保仿真阶段与实车部署阶段在信息流路径和特征表示上的一致性,为深度强化学习策略的稳定迁移与实时运行提供支撑。

图 14 RTRC 4Pro智能车部署平台

Fig. 14 Deployment platform of RTRC 4Pro intelligent vehicle

在实车验证中,进行了两项实验:

实验一:直线实验。其目标在于验证这两种模型在实际部署中的决策能力。该实验要求车辆在行驶过程中尽可能沿直线行驶,观察两种模型左右摇摆的情况。对每一种模型分别进行5次独立部署,每次部署均从相同的起点和姿态条件开始。当车辆在直线路段内未发生碰撞、未触碰两

侧挡板并完成既定距离行驶时，记为一次成功试验；若出现偏移过大导致出界或与挡板发生接触，则判定为失败。对于每一次部署，记录车辆沿直线行驶过程中的偏移量数据，并据此计算平均偏移距离；在此基础上统计5次部署中成功次数与成功率。通过比较平均偏移距离和成功率两个指标，可以定量评估两种模型在简单直线工况下的轨迹保持能力及其决策策略的稳健性，为后续整圈行驶实验的结果分析提供基线对照。

实验二：整圈行驶实验。本实验旨在评估模型在闭合回路上持续稳定行驶的能力。试验场地为一规则室内地图，由四段直线和四段弯道组成，其中单侧直线路段的有效长度为160 cm，弯道采用半径为60 cm的圆弧过渡，构成近似矩形的环形赛道，如图15所示。赛道两侧布置黑色挡板以模拟道路边界约束，车辆在行驶过程中不得与挡板发生碰撞或将其碰倒，一旦出现此类情况即判定为任务失败，从而重点考察模型对横向偏移和轨迹精度的控制能力。

图15　实验二场地全景图

Fig. 15　Panoramic view of the experiment 2 site

在上述地图条件下，实验二的具体测试流程如下：将训练完成的模型部署于实车平台，当车辆能够在不碰撞挡板的前提下连续完成五圈闭环行驶时，将该次部署视为一次有效实验。在每一次有效实验中，分别记录车辆每一圈的行驶时间和对应的偏移量序列，同时统计模型为达到“五圈连续成功”所需的总部署次数。通过对单圈时间、偏移量以及部署次数等指标的综合分析，可以系统评估不同模型在真实物理环境中的轨迹跟踪性能、运行稳定性及策略鲁棒性。

4.2　实车部署结果分析

实验一结果如图16所示，其中道路中心线由白色线条表示。LSTM-PPO模型(轨迹由红色点表示)表现较好。从轨迹图可以看出，红色点几乎完全沿着白色直线路径分布，横向偏差极小。这表明LSTM-PPO模型能够实现高度精确和稳定的运动控制，使车辆平滑且准确地跟踪预定路径。相比之下，PPO模型(轨迹由黄色点表示)表现则明显逊色。虽然在起点附近，黄色点与直线路径较为接近，但随着行驶距离增加，轨迹逐渐向右侧偏离，而后急速向左偏移，使轨迹形成弯曲的形状。

轨迹层面的差异在定量统计结果中得到了进一步印证。如表1所示，LSTM-PPO模型在直线实验中的平均偏移距离仅为2.268 cm，而PPO模型的平均偏移距离达到5.436 cm，前者约为后者的一半，说明基于时序信息增强后的策略能够显著抑制横向误差，使车辆在长距离直线行驶过程中始终保持在车道中心附近。同时，在5次部署中，LSTM-PPO模型实现了100%的成功率，而PPO模型的成功率仅为60%，表明在相同测试条件下，前者不仅轨迹更“直”，而且运行更加稳定可靠，更不容易因累积偏移而导致出界或碰撞。

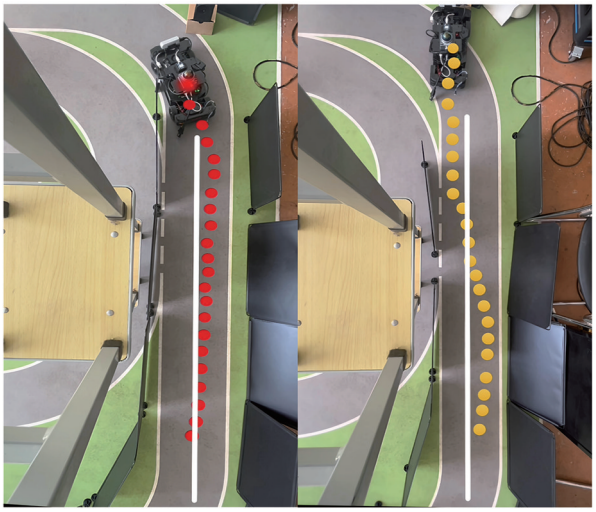


图16　直线行驶轨迹图

Fig. 16　Straight-line driving trajectory

表1　直线行驶实验统计结果

Tab. 1　Statistical results of the straight-line driving experiment

模型	平均偏移距离/cm	成功率/%
PPO	5.436	60
LSTM-PPO	2.268	100

实验二中，从整体结果来看，部署实验清晰地呈现出PPO模型与LSTM-PPO模型在轨迹跟踪性能上的差异。俯视轨迹图中(如图17所示)，PPO模型对应的黄色轨迹在弯道区域普遍向外鼓

出，特别是在长直道与弯道衔接的位置，采样点明显偏离理想行驶带，部分路段出现较大的横向漂移，说明该策略在应对曲率变化和连续控制时存在一定滞后与不稳定性。相比之下，LSTM-PPO模型的红色轨迹整体包络更加紧凑，轨迹点分布更加贴近赛道内侧边缘，弯道段的偏移范围显著收窄，表明在实际路面的场景中，引入时序记忆结构可以有效提高车辆的横向控制稳定性。

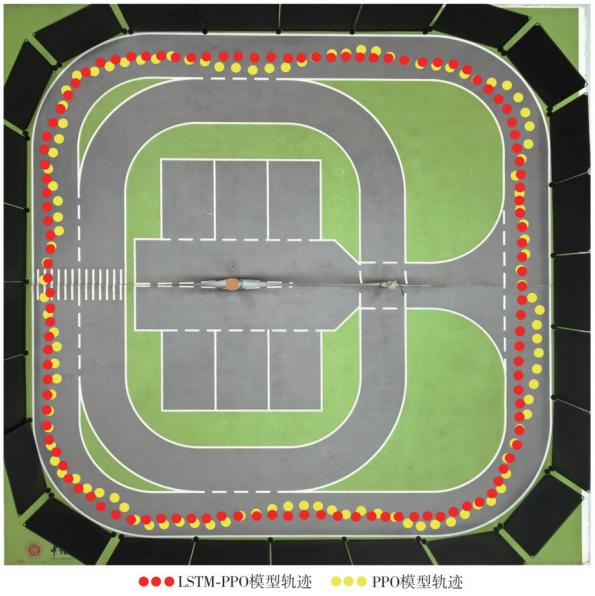
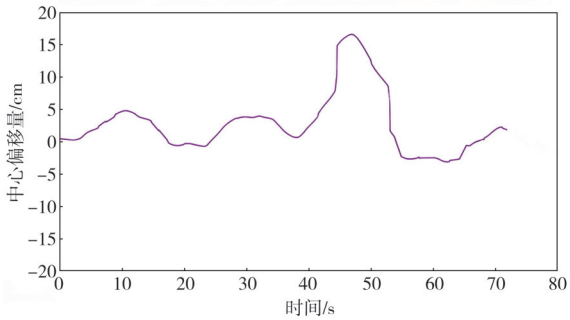


图 17 整圈行驶轨迹图

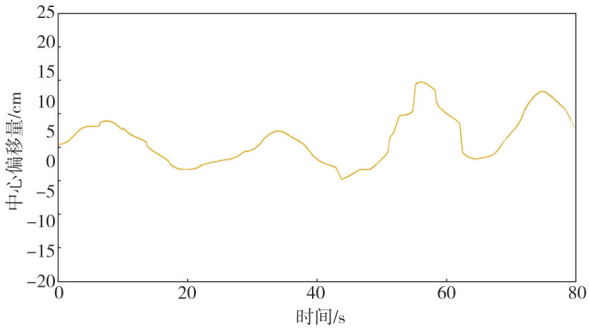
Fig. 17 Full-lap driving trajectories

这种空间分布上的直观差异在偏移量——时间折线图(如图 18 所示)中得到进一步验证。可以看到PPO模型的紫色曲线在整个单圈过程中起伏较为剧烈，尤其在大约中段时间区间内出现了明显的正向偏移峰值，随后又快速跌入负偏移区域，峰谷间的变化接近20 cm，反映出车辆在该区间经历了较大的横向摆动与频繁的修正动作。与之对应，LSTM-PPO模型的黄色曲线波动幅度明显减小，大部分时间被约束在较窄的偏移范围内，极端偏移事件显著减少，整体曲线呈现出更平稳、更“贴线”的形态，说明该模型能够更好地抑制误差积累和突发性偏移。

统计结果与上述时域和空间特征高度一致。如表 2 所示，偏移量的平均误差从 PPO 的 4.201 cm降低到LSTM-PPO的3.697cm，表明后者在整体偏移水平上更接近理想轨迹；均方根误差则由 19.868 cm 降至 7.356 cm，标准差也由 19.731 cm降至7.235 cm，二者下降均超过一半，说明LSTM-PPO不仅减小了普遍意义上的偏差，更重要的是显著抑制了偶发的大偏移和控制了抖动，使得车辆在整圈行驶过程中表现出更强的稳定性与可预测性。在单圈平均时间方面，PPO模型约为 76.37 s，而 LSTM-PPO 模型约为 84.65 s，后者用时略长，反映出其策略在纵向控制上相对更为保守，倾向于通过适度降低速度和减少激烈操纵来换取横向控制稳定和安全裕度的提升。这种“以时间换稳定”的决策在任务层面的表现良好，LSTM-PPO 的整体成功率约为62.50%，显著高于PPO的35.71%，在真实环境中达成完整闭环的概率提高了近一倍。



(a) PPO模型



(b) LSTM-PPO模型

图 18 平均偏移量

Fig. 18 Average offset

表 2 整圈行驶实验统计结果

Tab. 2 Statistical results of the full-lap driving experiment

模型	偏移量/cm			单圈平均时间/s	成功率/%
	平均误差	均方根误差	标准差		
PPO	4.201	19.868	19.731	76.37	35.71
LSTM-PPO	3.697	7.356	7.235	84.65	62.51

通过实车验证可知：LSTM-PPO通过引入长短期记忆网络(LSTM)，使得模型能够在连续动作空间中基于历史观测数据生成平滑且准确的决策输出，从而实现轨迹的高度贴合与极小的横向偏差。这一特性在动态环境中尤为关键，因为自动驾驶任务本质上涉及高维时序数据的处理，

LSTM的门控机制赋予了模型理解长期模式的能力。因此，LSTM-PPO模型的轨迹稳定平滑且实验二的失误更少。相比之下，PPO算法因缺乏对时序信息的深度挖掘，仅依赖当前状态进行决策，导致预测误差随时间累积，表现为轨迹的偏离与波动，实验二的失误次数较多。

5 结 论

本文针对深度强化学习在自动驾驶中面临的学习效率低、动作连续性不足及实际部署困难三大挑战，提出了一种融合预训练变分自编码器(VAE)与长短期记忆-近端策略优化(LSTM-PPO)的智能驾驶决策框架。通过利用VAE对高维语义图像进行特征降维，并与车辆状态信息融合，有效提升了模型的学习效率与收敛性；通过引入LSTM网络捕捉时序依赖，生成了更平滑、连贯的控制动作，显著改善了车辆的动作连续性；通过采用基于观测对齐与噪声注入的零样本迁移策略，实现了仿真模型到实车平台的无缝、鲁棒部署。仿真与实车实验结果一致表明，所提方法在轨迹跟踪精度、行驶稳定性与任务成功率上均显著优于传统PPO算法，验证了其在实际应用中的有效性与可靠性，为自动驾驶的深度强化学习工程化提供了有价值的实践路径。

利益冲突声明/Conflict of Interests

所有作者声明不存在利益冲突。

All authors declare no relevant conflict of interests.

参考文献：

[1] Ben Elallid B, Benamar N, Hafid A S, et al. A comprehensive survey on the application of deep and reinforcement learning approaches in autonomous driving [J]. Journal of King Saud University-Computer and Information Sciences, 2022, 34(9): 7366-7390.

[2] Kiran B R, Sobh I, Talpaert V, et al. Deep reinforcement learning for autonomous driving: a survey [J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(6): 4909-4926.

[3] 陈菁，赵聪，马裕城，等. 车路云一体化架构下面向舒适性提升的自动驾驶速度智能决策方法[J]. 中国公路学报，2025，38(2)：243-257.
Chen Jing, Zhao Cong, Ma Yucheng, et al. Intelligent speed planning approach for comfort improvement of autonomous vehicles based on vehicle-road- cloud integration architecture[J]. China Journal of Highway and Transport, 2025, 38(2): 243-257. (in Chinese)

[4] 方华珍，刘立，顾青，等. 基于换道意图和强化学习的车辆匝道汇入决策[J]. 无人系统技术，2024，7(5)：111-119.
Fang Huazhen, Liu Li, Gu Qing, et al. Vehicle on-ramp decision-making based on lane-change intention and reinforcement learning [J]. Unmanned Systems Technology, 2024, 7(5): 111-119. (in Chinese)

[5] 叶宝林，王欣，李灵犀，等. 基于深度强化学习PPO的车辆智能控制方法[J]. 计算机工程，2025，51(7)：385-396.
Ye Baolin, Wang Xin, Li Lingxi, et al. Vehicle intelligent control method based on deep reinforcement learning PPO[J]. Computer Engineering, 2025, 51(7): 385-396. (in Chinese)

[6] 宋建融，刘丽. 智能网联新能源汽车自动驾驶控制算法研究[J]. 汽车测试报告，2024(22)：68-70.

[7] Zhang D, Liang J, Guo K, et al. CarPlanner: consistent auto-regressive trajectory planning for large-scale reinforcement learning in autonomous driving [C]// 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025: 17239-17248.

[8] Zhao Y N, Wu K, Xu Z Y, et al. CADRE: a cascade deep reinforcement learning framework for vision-based autonomous urban driving [C]//AAAI Conference on Artificial Intelligence, 2022: 20259.

[9] Hafner D, Pasukonis J, Ba J, et al. Mastering diverse domains through world models[PP/OL]. arXiv (2023-01-10) [2025-07-23]. https://arxiv. org/pdf/2301.04104v1.

[10] Chen J, Li S E, Tomizuka M. Interpretable end-to-end urban autonomous driving with latent deep reinforcement learning [J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(6): 5068-5078.

[11] Chen Y, Ji C, Cai Y, et al. Deep reinforcement learning in autonomous car path planning and control: a survey [PP/OL]. V2. arXiv (2024-03-30) [2025-07-23]. https://doi. org/10. 48550/arXiv. 2404. 00340.

[12] Wu J, Huang C, Huang H, et al. Recent advances in reinforcement learning-based autonomous driving behavior planning: a survey [J]. Transportation Research Part C: Emerging Technologies, 2024, 164: 104654.

[13] Zhang Y, Zhao W, Wang J, et al. Recent progress, challenges and future prospects of applied deep reinforcement learning: a practical perspective in path planning[J]. Neurocomputing, 2024, 608: 128423.

[14] 周昕阳, 宋振波, 李蔚清, 等. 基于LSTM深度强化学习的端到端自动驾驶[J]. 计算机仿真, 2024, 41(2): 172-178.
Zhou Xinyang, Song Zhenbo, Li Weiqing, et al. End-to-end autonomous driving based on LSTM deep reinforcement learning combined [J]. Computer Simulation, 2024, 41(2): 172-178. (in Chinese)

[15] Zhao J W, Qu T, Xu F. A deep reinforcement learning approach for autonomous highway driving [J]. IFAC-PapersOnLine, 2020, 53(5): 542-546.

[16] Jagadish D N, Chauhan A, Mahto L. Conditional variational autoencoder networks for autonomous vehicle path prediction[J]. Neural Processing Letters, 2022, 54(5): 3965-3978.

[17] Oh G, Peng H. CVAE-H: conditionalizing variational autoencoders *via* hypernetworks and trajectory forecasting for autonomous driving[PP/OL]. arXiv (2022-01-24) [2025-07-23]. https://doi. org/10. 48550/arXiv. 2201. 09874.

[18] Wang S F, Wang Z L, Wang X K, et al. Intelligent vehicle driving decision-making model based on variational AutoEncoder network and deep reinforcement learning[J]. Expert Systems with Applications, 2025, 268: 126319.

[19] Lai Z H, BräUnl T. End-to-end learning with memory models for complex autonomous driving tasks in indoor environments [J]. Journal of Intelligent & Robotic Systems, 2023, 107(3): 37.

[20] 丁云龙, 匡敏驰, 朱纪洪, 等. 基于LSTM-PPO算法的多机空战智能决策及目标分配[J]. 工程科学学报, 2024, 46(7): 1179-1186.
Ding Yunlong, Kuang Minchi, Zhu Jihong, et al. Intelligent decision making and target assignment of multi-aircraft air combat based on the LSTM-PPO algorithm [J]. Chinese Journal of Engineering, 2024, 46(7): 1179-1186. (in Chinese)